MEDLINE®/PubMed® Baseline Repository (MBR) Reference Material

Date Last Updated: Friday, September 02, 2005

1. Introduction

Researchers from time to time have requested the ability to have available MEDLINE citations in the state they were at a given moment in time without the yearly updates and continual revisions that occur. The MEDLINE/PubMed Baseline Repository (MBR) was setup to provide this capability. We have stored the end of year baseline of the MEDLINE/PubMed database for each year starting in 2002 along with a selection of the associated MeSH Vocabulary data files.

We have created tools to do some basic calculations like frequency counts of MeSH terms for each of the baselines. In addition, we have developed a database for each baseline which allows users to perform queries on specific baselines and retrieve subsets pertinent to their research quickly and easily.

2. Restrictions/Caveats

Users are responsible for compliance with all applicable MEDLINE/PubMed and other NLM® Databases License Agreements.

The records included in the MEDLINE/PubMed Baseline databases represent a static view of the data at the time each baseline database was created.

To access the MeSH files, you must enter into an online Memorandum of Understanding for use of the MeSH Vocabulary data.

To access the MBR Query Tool, you must be a recognized licensee of NLM Data. The License agreement requires those who use the MEDLINE/PubMed database to fill-out an Intended Use Worksheet once a year and to file a brief Usage Report (Note: research use licensees do not have to submit the Usage Report form) to summarize their use of the database. NLM leases MEDLINE/PubMed to U.S. individuals or organizations; to its formally recognized International MEDLARS Centers; and to non U.S. individuals or organizations for internal research projects with no commercial citation search service.

3. Baselines

The baselines are normally generated towards the middle of November each year and contain all ***completed*** citations in MEDLINE as of that date. The baselines represent MEDLINE ***after*** the year-end processing has been completed. This means that the records have been revised with the upcoming year's new MeSH vocabulary terms. We currently have available the 2002, 2003, 2004, and 2005 MEDLINE/PubMed Baselines. The naming of the baselines represents this year-end processing. For example, the 2002 MEDLINE/PubMed Baseline contains all ***completed*** citations from the mid-1960's until the date the baseline was created in late November 2001 with the year-end processing assigning appropriate 2002 MeSH vocabulary terms, thus it is a baseline for the 2002 year.

The baselines contain citations that are not MEDLINE as well. All of the baselines we have stored (2002 on) contain "Out-of-scope" citations which were renamed to "PubMed-not-MEDLINE" starting with the 2004 MEDLINE/PubMed Baseline. The PubMed-not-MEDLINE status refers to citations that reside in PubMed from journals included in MEDLINE and have undergone quality review but are not assigned MeSH headings because the cited item is not in scope for MEDLINE either by topic or by date of publication. Citations in the Out-of-scope or PubMed-not-MEDLINE status make up a very small percentage (0.51% or 75,271 records in the 2005 baseline) of the total number of citations contained in the baselines.

Starting with the 2005 MEDLINE/PubMed Baseline, OLDMEDLINE citations are also included in the baselines. The OLDMEDLINE citations make up 11.9% (or 1,760,574) of the total number of citations contained in the 2005 baseline. The OLDMEDLINE citations are from international biomedical journals covering the fields of medicine, preclinical sciences, and allied health sciences. The citations were originally printed in hardcopy indexes published prior to 1966. For additional information, please refer to the following URL: http://www.nlm.nih.gov/databases/databases_oldmedline.html. Currently, the subject indexing from the OLDMEDLINE citations is being stored in the "Other Term" (or "OT") tagged fields and not the MeSH Terms (or MH) tagged fields. This means that searching from our MBR Query Tool via the MH field will not include any OLDMEDLINE citations. The only way to include OLDMEDLINE records will be to do a timeframe query without specifying any field specific search criteria.

Baseline	Created	Number of Citations
2002	around November 21, 2001	11,299,108
2003	between November 1-4, 2002	11,847,524
2004	between November 14-18, 2003	12,421,396
2005	November 20, 2004	14,792,864

List of Available Baselines

MeSH Files: We have preserved the following list of “MeSH in ASCII format” and “MeSH in XML format” files for each of the baseline years represented in the Repository. The MeSH files are accessible upon completion of the MeSH Memorandum of Understanding.

File	Description
cYYYY.bin **suppYYYY.xml**	Supplementary Concept Records (formerly Supplementary Chemical Records), Data Elements file.
dYYYY.bin **descYYYY.xml**	Descriptor Data Elements file.
mtreesYYYY.bin	MeSH main headings with the tree numbers that place the heading in a hierarchical arrangement. Sorted by tree number. ASCII format.
qYYYY.bin **qualYYYY.xml**	Qualifier Data Elements

We have also developed the "***MEDLINE/PubMed Baseline Repository (MBR)***" website which may be accessed at the following URL: http://mbr.nlm.nih.gov. The website allows public access to all of the reference materials and resources we have for each baseline, while maintaining access control on the MBR Query Tool. Access to the MBR Query Tool is regulated via the IP addresses associated with each NLM recognized licensee of MEDLINE/PubMed data. The website includes the following:

- The "***MEDLINE/PubMed Baseline Query Tool***", a web-based tool allowing users to perform baseline specific queries and to create subsets and test collections.
- A download page where users can download all of the MBR files we discuss in the table below.
- A web page which provides a detailed set of reference material about the Repository.

For each of the Baselines, we have provided the following resources available from our website (see above):

Resource	Restrictions
MBR Query Tool Database: Baseline databases 2002 forward available for searching. Includes tables with MH, SH, MH/SH combination, Chemicals, and PMID data; also can limit or filter by Date Created, Date Completed, Date Last Revised, Publication Year, and Status.	License Required
Orignal XML Formatted Citations: Original XML version of baseline citations.	License Required
MEDLINE ASCII Display Formatted Citations: Each XML citation translated to MEDLINE ASCII display format used in PubMed.	License Required
DTD Files: We save a copy of the relevant DTD (Document Type Definition) files each year for working with the Baseline XML files.	No Restrictions
Frequency Count Files: Basic frequency counts for the entire MEDLINE/PubMed Baseline sorted into alphabetical and numerical order for the following MEDLINE fields. For all fields but the NM field, we also provide a sort and count of their occurrences as starred (Index Medicus) items. a. MH (MeSH Headings) b. SH (MeSH Subheadings) c. MH/SH combinations d. NM (Chemicals)	No Restrictions
Raw Data Files: Files containing the raw data similar to what was used to create our MBR Query Tool Database for this Baseline year. There is a README file describing the various files available and their layouts.	No Restrictions
Histogram/Summary Files: File showing the number of MH terms assigned to each of the various MeSH Tree top-level and top-level + 1 categories during the latest year to see how assignment of terms might vary from year to year. File showing the number of MH terms assigned to each of the UMLS Semantic Type Groupings categories during the latest year to see how assignment of terms might vary from year to year from a different perspective.	No Restrictions
Related MeSH Files: We save a copy of selected MeSH Vocabulary data files for each year and a copy of their associated DTD (Document Type Definition) files for working with the Baseline XML files.	Memorandum of Understanding Required
UMLS Semantic Groups File: We have saved a copy of the Semantic Groups file. The Semantic Groups are a coarse-grained set of semantic type groupings designed to reduce the complexity in the UMLS Metathesaurus. The 15 semantic groups provide a partition of the UMLS Metathesaurus for 99.5% of the concepts.	No Restrictions

4. Frequency Counts:

Simple frequency counts were done of all MH and RN lines for all citations in each of the baseline collections. These frequency counts were then aggregated into the following categories: Chemical (RN terms), Main Heading (MH terms), SubHeading (qualifier terms), and Main Heading/SubHeading (MH/qualifier combination terms). The results are provided either in alphabetical or numerical sorted order for each category.

Chemical Files (RN terms):

-- Format of files:

Frequency count overall in MEDLINE
Registry Number for the Chemical
Chemical name

-- Chemical_freq_alpha -- Ordered by Chemical name
-- Chemical_freq_count -- Ordered by Frequency counts (high->low)

Main Heading (MH) Files (MH terms):

-- Format of files:

Frequency count overall in MEDLINE
Frequency count when starred (major) item
MeSH Heading

-- MH_freq_alpha -- Ordered by MeSH Headings
-- MH_freq_count -- Ordered by Overall Frequency counts (high->low)
-- MH_major_freq_count - Ordered by Starred Frequency counts (high->low)

SubHeading (SH) Files (qualifier terms):

-- Format of files:

Frequency count overall in MEDLINE
Frequency count when starred (major) item
SubHeading or Qualifier name

-- SH_freq_alpha -- Ordered by Subheadings/qualifier
-- SH_freq_count -- Ordered by Overall Frequency counts (high->low)
-- SH_major_freq_count - Ordered by Starred Frequency counts (high->low)

Main Heading/SubHeading (MH_SH) Combination Files (MH/qualifier combination terms):

-- Format of files:

Frequency count overall in MEDLINE
Frequency count when starred (major) item
MH/SubHeading name combination

-- MH_SH_freq_alpha -- Ordered by MH/SH
-- MH_SH_freq_count -- Ordered by Overall Frequency counts (high->low)
-- MH_SH_major_freq_count - Ordered by Starred Frequency counts (high->low)

Special Note: This counts each of the MH/SH combinations, applying the starred count only to combinations that include the starred term - for example:

If we have an entry like the following in a citation:

MH - Obesity/*complications/etiology

We include the following entries in this table:

Obesity/*complications
Obesity/etiology

5. MH Term Assignment by MeSH Treecodes Summary (hist)

The MH Term assignment summary was done using the information garnered from the frequency counts. We found the MeSH Tree code for each of the unique MH terms found for each baseline and then summarized the frequency counts for the top two MeSH Tree levels. The idea for this summary was to see if we could detect any patterns in the way MeSH terms are assigned from year to year by creating histograms comparing the various years either within a baseline or across baselines. The example below is a view from the 2004 MEDLINE Baseline summary file.

Top-Level Summary:

```
Anatomy [A]|357747
Organisms [B]|392381
Diseases [C]|537255
Chemicals and Drugs [D]|980930
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]|877038
Psychiatry and Psychology [F]|185296
Biological Sciences [G]|1024627
Physical Sciences [H]|383770
Anthropology, Education, Sociology and Social Phenomena [I]|109299
Technology and Food and Beverages [J]|45889
Humanities [K]|27983
Information Science [L]|149576
Persons [M]|388012
Health Care [N]|529399
Geographic Locations [Z]|94882
```

Second-Level Summary:

```
Body Regions [A01]|14835
Musculoskeletal System [A02]|32999
Digestive System [A03]|20192
Respiratory System [A04]|8780
Urogenital System [A05]|18956
Endocrine System [A06]|8560
Cardiovascular System [A07]|30369
Nervous System [A08]|63992
Sense Organs [A09]|11155
Tissues [A10]|36849
…
Population Characteristics [N01]|59665
Health Care Facilities, Manpower, and Services [N02]|75251
Health Care Economics and Organizations [N03]|59176
Health Services Administration [N04]|106004
Health Care Quality, Access, and Evaluation [N05]|322311
Geographic Locations [Z01]|94882
```

6. MH Term Assignment by UMLS Semantic Type Group Summary (histST)

The MH Term assignment summary was done using the information garnered from the frequency counts. We found the UMLS Semantic Type(s) for each of the unique MH terms found for each baseline and then summarized the frequency counts for each of the UMLS Semantic Type Groups which are high-level categories combining multiple Semantic Types. The idea for this summary was another way to see if we could detect any patterns in the way MeSH terms are assigned from year to year by creating histograms comparing the various years either within a baseline or across baselines. The example below is a view from the 2004 MEDLINE Baseline summary file.

```
Activities & Behaviors|164011
Anatomy|364259
Chemicals & Drugs|1719653
Concepts & Ideas|642821
Devices|30862
Disorders|640501
Genes & Molecular Sequences|76533
Geographic Areas|95287
Living Beings|798945
Objects|65613
Occupations|71534
Organizations|29226
Phenomena|128390
Physiology|408765
Procedures|747397
```

7. Histogram/Summary Files

We have created a couple of histograms or summaries of how the MeSH Headings (MH) are assigned. The first perspective is from how the MHs are assigned based on their MeSH Vocabulary Treecodes. In the MeSH Vocabulary, each MH is assigned one or more treecodes, and we assign counts for all of the treecodes assigned. The second perspective is how the MHs are assigned based on their UMLS Semantic Types. Each UMLS concept (MHs are concepts in UMLS) is assigned one or more Semantic Types and we assign counts for all of the Semantic Types assigned. We then roll the individual counts for the Semantic Types up into the UMLS Semantic Groupings. We have also taken the time to graph the various histograms to provide a visual as well as data driven view of the data.

The Treecodes and Semantic Types help describe the MeSH Headings and place them appropriately within either MeSH or the UMLS. For more information on MeSH Treecodes, UMLS Semantic Types, and UMLS Semantic Groupings, please refer to the "References" section at the bottom of this page.

File	Description
hist	A count of MeSH Main Headings partitioned into the respective MeSH Treecodes for this year. We include counts for the top-level (e.g., "A", "B") and top-level plus one (e.g., "A01", "A02, "B01") MeSH Treecodes based on the MeSH Vocabulary specific to each baseline year.
hist.pdf	A graph generated from the respective hist file data.
histST	A count of UMLS Semantic Types partitioned into the respective high-level UMLS Semantic Groupings.
histST.pdf	A graph generated from the respective histST file data.
hist_Full	We use the same methodology we used in the above "hist" file, but, apply it across the entire MEDLINE and report counts for each of the years (1965 - 2003) in the MEDLINE Baseline. This file is provided for comparison purposes to see how assignments have changed over the years. **NOTE:** This file is created using the 2004 MeSH Vocabulary, so the numbers will differ from the 2002 and 2003 baseline counts.
hist_Full.pdf	A graph generated from the respective hist_Full file data.
histST_Full	We use the same methodology we used in the above "histST" file, but, apply it across the entire MEDLINE and report counts for each of the years (1965 - 2003) in the MEDLINE Baseline. This file is provided for comparison purposes to see how assignments have changed over the years. **NOTE:** This file is created using the 2004 MeSH Vocabulary, so the numbers will differ from the 2002 and 2003 baseline counts.
histST_Full.pdf	A graph generated from the respective histST_Full file data.
combined_hist.pdf	A graph generated from the "hist" methodology results for all of the MEDLINE Baselines currently in the Repository for comparison purposes.
combined_histST.pdf	A graph generated from the "histST" methodology results for all of the MEDLINE Baselines currently in the Repository for comparison purposes.

NOTES:

1. All counts are based solely on MeSH Headings.
2. We use the Date Completed date to determine inclusion in the counts for each baseline.
3. We are using the MeSH assignment of Semantic Type for the MeSH Headings wherever possible. Where a Semantic Type was not found in the MeSH dYYYY.bin file for a given MeSH Heading, we manually determined the Semantic Type(s) via the UMLS MRCON and MRSTY files.
4. The hist_Full and histST_Full files which report on the entire MEDLINE baseline were created using the latest MeSH Vocabulary, 2004 in this case. This means that some counts may be different in this overall file than what is seen in the 2002 and 2003 baseline count files. The 2004 MeSH was used to provide consistency with using the 2004 Baseline to create the two files.

8. MySQL® Database

A MySQL database is created for each of the MEDLINE Baselines that we have. The purpose of the databases is to provide a search capability to allow researchers to query a given baseline year easily. This capability allows for easy creation of custom test collections and provides an easy means to focus research efforts. The following tables are included in each baseline database:

Data Tables

Chemical_data:

Field	Type
SubstanceName	tinytext
PMID	int(11)

MH_data:

Field	Type
Term	tinytext
MajorTopic	smallint(6)
PMID	int(11)

SH_data:

Field	Type
Term	tinytext
MajorTopic	smallint(6)
PMID	int(11)

MH_SH_data:

Field	Type
Terms	tinytext
MajorTopic	smallint(6)
PMID	int(11)

PMID_data:

Field	Type
PMID	int(11)
DateCreated	date
DateCompleted	date
DateRevised	date
PubYear	year(4)
Status	tinyint(4)

Lists used for Field Validation

MH_list:

Field	Type
Term	varchar(104)
count	int(11)

MH_SH_list:

Field	Type
Term	varchar(129)
count	int(11)

SH_list:

Field	Type
Term	varchar(30)
count	int(11)

RN_list:

Field	Type
Term	varchar(245)
count	int(11)

NOTE on SH: There are periodically duplicate entries of SHs in a given citation. We may also have a mixture of starred and non-starred entries with the same SH. We attempt in this table to capture the uniqueness of the SH within the citation.

For example (PMID: 7176534):

MH - Complement 3/*analysis
MH - Immunoglobulin A/analysis
MH - Immunoglobulin G/analysis
MH - Immunoglobulin M/analysis
MH - Immunoglobulins/*analysis

We would add the following entries into the SH table which shows that for this SH in this citation, we have a uniqueness of "analysis" appearing as a starred item AND as a non-starred item:

analysis|0|7176534
analysis|1|7176534

Description of the Fields:

term(s) - MeSH Main Heading, MeSH SubHeading/Qualifier, or a combination of these two entities.

MajorTopic - Flag as to whether term(s) are starred. 1 = yes, 0 = no.

PMID - PubMed Identifier

DateCreated - Date that the citation was recognized as being in PubMed. ***NOTE Format:*** YYYYMMDD

DateCompleted - Date that record was recognized as being completed. ***NOTE Format:*** YYYYMMDD

DateRevised - Date that the citation was revised. ***NOTE Format:*** YYYYMMDD

PubYear – Year citation was published - from the Publication Date field. ***NOTE Format:*** YYYY

Status – Status of the citation.
0 – MEDLINE
1 – OLDMEDLINE
2 – PubMed-not-MEDLINE/Out-of-scope

SubstanceName - Name of the Substance

9. MEDLINE Baseline Repository Query Tool

The MEDLINE Baseline Repository Query Tool allows users to query any of the currently stored MEDLINE Baselines (2002, 2003, 2004, and 2005). Users may use any combination of MeSH Headings (MH), MeSH SubHeadings (SH), Supplemental Concepts/Chemicals (RN), and MH/SH combinations to build their query searches. Users may also specify a date range to help limit the search.

The results of the query search can be provided either as a list of matching PMIDs or as the matching citations themselves. If the user chooses to receive the matching citations, they can select between XML or MEDLINE formats. The user may also elect to have the results (either PMID list or citations) randomized and split into Testing and Training subsets.

The ability to create subsets was designed to provide researchers a simple way to develop test collections for their work. Depending on the researcher's needs, we have the capability to create a single subset of the requested size, or we can automatically split the results into Testing and Training subsets based on the researcher's requested distribution. If requested, we can also randomize the ordering of the results before we create the subsets.

Using the Query Tool, a researcher could very easily develop a test collection based on the following scenario:

> Using the 2003 MEDLINE Baseline, I would like all citations in Full MEDLINE format that include the MeSH Heading "Liver" and the MeSH SubHeading "drug therapy" (not necessarily in "Liver/drug therapy" combination) that were completed (using Date Completed/DCOM date) between January 1, 2002 and August 31, 2002. I need to have a randomized test collection that includes Testing and Training subsets with 90% of the results placed in the Training subset and the remaining 10% in the Testing subset.

10. MEDLINE/PubMed Baseline Repository Detailed Resources Information

This section of the paper details where, when, how, and what resources are used for creating and maintaining the MEDLINE/PubMed Baseline Repository.

MEDLINE/PubMed Baseline:

The MEDLINE/PubMed Baselines are generated each year by the U.S. National Library of Medicine (NLM). The baseline is typically generated towards the end of November each year and officially announced around the middle of December each year. For status and information pertaining to the MEDLINE/PubMed Baselines, you can look at the Bibliographic Services Division (BSD) "Information for Licensees of NLM Data" web site located at http://www.nlm.nih.gov/bsd/licensee.html. Here you will find information in the "MEDLINE Documentation" and "Announcements and General Information" sections of the page. NLM's recognized licensee's of MEDLINE/PubMed data have access to the Baselines via the MBR Query Tool.

Year Specific MeSH Vocabulary Files:

We also maintain a copy of each year's specific MeSH Vocabulary files so that the appropriate MeSH information is available for each baseline. It's important to remember that the MeSH Vocabularies follow the MEDLINE/PubMed Baseline yearly naming convention because the year-end processing involves using the new MeSH Vocabulary. For example, the 2004 MEDLINE/PubMed Baseline created in November of 2003 uses the 2004 MeSH Vocabulary.

To retrieve the MeSH Vocabulary files for each year, we go to the Medical Subject Headings (MeSH) download web site at http://www.nlm.nih.gov/mesh/filelist.html, agree to the Memorandum of Understanding, fill in the data request form, and then ftp down all of the available ASCII formatted files which currently includes: Descriptor Data Elements (dYYYY.bin), Supplementary Concept Records (cYYYY.bin), Qualifier Data Elements (qYYYY.bin), and the YYYY MeSH Trees (mtreesYYYY.bin) file. This web site fortunately preserves at least the prior year's Vocabulary as well as the current one.

UMLS Semantic Types Grouping (SemGroups.txt) file:

The SemGroups.txt file is the latest addition to the Repository and it's unclear whether this file is updated each year, as the Semantic Types change, or is static. This file has grouped the UMLS Semantic Types into 15 (currently) high-level categories. We are using this file to see if we can detect patterns in how the MeSH Headings are assigned in MEDLINE. The papers: "***Aggregating UMLS semantic types for reducing conceptual complexity*.** McCray AT, Burgun A, Bodenreider O; Medinfo. 2001;10(Pt 1):216-20." and "***Exploring semantic groups through visual approaches*.**, Bodenreider O, McCray AT; Journal of Biomedical Informatics. 2003; 36(6):414-432." provide much greater detail on the grouping of the Semantic Types. Both papers can be found at the Lister Hill National Center for Biomedical Communications web site (http://lhncbc.nlm.nih.gov).

To retrieve the SemGroups.txt file, we go to the UMLSKS (Unified Medical Language System Knowledge Source Server) at http://umlsks.nlm.nih.gov, login, and then go to the "Semantic Groups" link under the "UMLSKS Resources" section of the welcome page.